5-2016

Discovery of a new repeat family in the Callithrix jacchus genome

Miriam Konkel
Louisiana State University and Agricultural and Mechanical College, konkel@lsu.edu

Mark Batzer
Louisiana State University and Agricultural and Mechanical College, mbatzer@lsu.edu

Brygg Ullmer
Louisiana State University and Agricultural and Mechanical College, ullmer@lsu.edu

Erika L. Arceneaux

Sreeja Sanampudi

See next page for additional authors

Discovery of a new repeat family in the *Callithrix jacchus* genome

Miriam K. Konkel,[1,4] Brygg Ullmer,[2,4] Erika L. Arceneaux,[1] Sreeja Sanampudi,[1] Sarah A. Brantley,[1] Robert Hubley,[3] Arian F.A. Smit,[3] and Mark A. Batzer[1]

[1] *Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA;* [2] *School of Electrical Engineering and Computer Science, Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana 70803, USA;* [3] *Institute for Systems Biology, Seattle, Washington 98109-5263, USA*

We identified a novel repeat family, termed Platy-l, in the *Callithrix jacchus* (common marmoset) genome that arose around the time of the divergence of platyrrhines and catarrhines and established itself as a repeat family in New World monkeys (NWMs). A full-length Platy-l element is ~100 bp in length, making it the shortest known short interspersed element (SINE) in primates, and harbors features characteristic of non-LTR retrotransposons. We identified 2268 full-length Platy-l elements across 62 subfamilies in the common marmoset genome. Our subfamily reconstruction and phylogenetic analyses support Platy-l propagation throughout the evolution of NWMs in the lineage leading to *C. jacchus*. Platy-l appears to have reached its amplification peak in the common ancestor of current day marmosets and has since moderately declined. However, identification of more than 200 Platy-l elements identical to their respective consensus sequence, and the presence of polymorphic elements within common marmoset populations, suggests ongoing retrotransposition activity. Platy-l, a SINE, appears to have originated from an *Alu* element, and hence is likely derived from 7SL RNA. Our analyses illustrate the birth of a new repeat family and its propagation dynamics in the lineage leading to the common marmoset over the last 40 million years.

[Supplemental material is available for this article.]

The common marmoset (*Callithrix jacchus*), also known as the white-tufted-ear marmoset, is a platyrrhine native to the Atlantic coastal forest of northeastern Brazil. Platyrrhines, commonly referred to as New World monkeys (NWMs), are primates indigenous to South America. They represent a diverse group of animals that diverged from catarrhines about 35–47 million years ago (Schrago and Russo 2003; Perelman et al. 2011; The Marmoset Genome Sequencing and Analysis Consortium 2014). The overall repeat content of the *C. jacchus* genome, which represents the first sequenced NWM genome, is similar to other previously analyzed primate genomes, with transposable elements making up at least half the genome mass (International Human Genome Sequencing Consortium 2001; The Chimpanzee Sequencing and Analysis Consortium 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007; Locke et al. 2011; Carbone et al. 2014; The Marmoset Genome Sequencing and Analysis Consortium 2014).

The major drivers of repeat-driven genome expansion in primate genomes are retrotransposons. They move in a "copy-and-paste" fashion throughout the genome using an RNA intermediate (Ostertag and Kazazian 2001; Belancio et al. 2008; Cordaux et al. 2009) and can be subdivided into elements with long terminal repeats (LTRs), i.e., endogenous retroviruses, and those that lack LTRs. The latter are commonly referred to as non-LTR retrotransposons. In primates, the two major non-LTR elements and the largest contributors to genome expansion are long interspersed element 1, L1 (LINE1) and the primate-specific *Alu* element, a short interspersed element (SINE) (International Human Genome Sequencing Consortium 2001; Cordaux et al. 2009; Konkel et al. 2010). A full-length *Alu* element is ~300 bp in length and terminates in an adenosine-rich tail (A-tail), a typical characteristic of non-LTR retrotransposons (Batzer and Deininger 2002). L1 predates the origin of primates, has been active throughout the radiation of primates, and represents the only known currently propagating autonomous non-LTR retrotransposon in primate genomes (Smit 1999; Ostertag and Kazazian 2001; Cordaux et al. 2009; Burns and Boeke 2012; Huang et al. 2012). LINEs and SINEs propagate in genomes via a process termed target-primed reverse transcription (TPRT) (Luan et al. 1993; Dewannieux et al. 2003) by utilizing the reverse transcriptase and endonuclease encoded by open reading frame 2 (ORF2) (Mathias et al. 1991; Feng et al. 1996; Cordaux et al. 2009). As a consequence of TPRT, classical retrotransposon insertions share hallmarks such as target site duplications (TSDs) (International Human Genome Sequencing Consortium 2001; Szak et al. 2002; Cordaux et al. 2009).

Following the divergence of two primate lineages, each lineage evolves uniquely and independently. Over time, this results in distinctive changes specific to each lineage; mobile elements are no different in this regard. Consequently, each lineage accumulates lineage-specific mobile element insertions and mobile element-mediated rearrangements (Cordaux et al. 2009; Konkel and Batzer 2010; Konkel et al. 2010). Moreover, mobile element subfamilies evolve uniquely in each lineage. In this respect, the lineage leading to the common marmoset had at least 35 million years (Perelman et al. 2011) of platyrrhine-specific evolution

[4] **These authors contributed equally to this work.**
Corresponding authors: konkel@lsu.edu, mbatzer@lsu.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.199075.115.

(i.e., not shared with catarrhines) leading to thousands of lineage-specific mobile element insertions primarily generated by NWM-specific subfamilies (The Marmoset Genome Sequencing and Analysis Consortium 2014).

A relatively small number of source elements are responsible for the bulk of non-LTR retrotransposon insertions (Deininger et al. 1992; Batzer and Deininger 2002; Brouha et al. 2003; Han et al. 2005; Walker et al. 2012), meaning the majority of insertions are dead upon arrival, i.e., they do not generate daughter copies. Moreover, the stealth model of propagation proposes that a small number of retrotransposition-competent elements propagate at a slow rate over extended periods of time (Han et al. 2005). Some daughter elements are highly active and generate many insertions in a relatively short time. These elements are likely deleterious to the host and, thus, are often lost from the population relatively quickly.

Results

We identified a nucleotide sequence of unknown origin specific to *C. jacchus* on Chr 3 (139597231–139598086) based on a multiple sequence alignment of common marmoset (calJac3.2), human (hg19), chimpanzee (panTro2/4), orangutan (ponAbe2), rhesus macaque (rheMac2/3), and squirrel monkey (saiBol1) that was confirmed by locus-specific PCR using a phylogenetic panel (Supplemental Table S1A; Supplemental Fig. S1). Our RepeatMasker (Smit et al. 2013–2015) analysis revealed that part of the sequence of unknown origin was not identified as a repeat. Based on manual inspection, we determined that this sequence was ~100 bp in length and followed by an adenosine-rich tail (hereafter referred to as A-tail). In subsequent analyses using BLAT (Kent 2002) and BLASTN (Altschul et al. 1990), we determined that the original sequence of unknown origin was not recognized in the human, chimpanzee, orangutan, or rhesus macaque genome assemblies. In contrast, many high homology matches were identified in the common marmoset genome assembly, indicating identification of a novel repeat family, which we termed Platy-1 because of its discovery in a NWM species and its apparent limited distribution in platyrrhines.

Characterization of Platy-l elements

Our RepeatMasker (Smit et al. 2013–2015) analysis of the *C. jacchus* assembly using a custom library (see Supplemental Methods S1) retrieved 2183 full-length Platy-1 elements. Based on 474 elements from six chromosomes (Chr 3, 10–13, and 17) (Supplemental Table S2), we determined the basic structure of the element (Fig. 1A) and investigated typical sequence features. About 90% of the insertions contained TSDs of ≥7 bp, with a length spectrum resembling a Poisson distribution with a median and average length of 15 bp (Fig. 1B), similar to TSD lengths of other non-LTR retrotransposons (Roy-Engel et al. 2002; Szak et al. 2002; Chen et al. 2005; Zingler et al. 2005). Further analysis of all Platy-1 elements with TSDs confirmed the presence of an endonuclease cleavage site (Fig. 1D) and termination in an A-tail of varying length without a polyadenylation signal.

The majority of Platy-1 elements have an A-tail between 11 and 20 nucleotides, with 8.2% (57/427) having a pristine A-tail larger than 20 bp (Fig. 1C). About one-third (37.6%) of the 427 analyzed putative full-length Platy-1 elements have A-tails harboring nucleotides other than adenosine. The two longest A-tails of putative Platy-1 elements contained tetranucleotide microsatellites of 13 and 15 repeat units, respectively. Approximately 55% (41 of 75) of Platy-1 loci with an A-tail larger than 20 bp and at most one substitution harbored a Pol III termination signal within 100 bp of the downstream flanking sequence and 21 loci within 25 bp. Some of these loci may be source elements and, thus, may have contributed to the expansion of the Platy-1 repeat family.

Platy-l subfamily reconstruction and evolution

Our Platy-1 subfamily structure analysis of all 2183 full-length Platy-1 elements, for which we used a majority rule approach, retrieved 62 Platy-1 subfamilies and 2275 elements (Supplemental Methods S1/Material S6; Supplemental Fig. S2). The higher number of Platy-1 elements in our final data set is a direct result of the subfamily reconstruction, as elements too diverged to be initially identified as a Platy-1 element were discovered with a more defined subfamily library. Of the 209 Platy-1 elements with an asterisk in our RepeatMasker output, indicating a second, higher scoring match with partial overlap (Smit et al. 2013–2015), we determined that the vast majority (95%) represent true Platy-1 insertions (Supplemental Methods S1). The remaining seven loci appear to be derived from a different repeat family. Thus, we estimate that the common marmoset genome contains 2268 full-length Platy-1 elements (Supplemental Table S3) translating to a density of approximately 0.77 Platy-1 elements per megabase.

As expected, the Platy-1 subfamily sizes vary (Table 1), with the largest subfamily (Platy-1-13) including 160 members and the smallest subfamily (Platy-1-16c) encompassing seven elements. More than two-thirds (43/62) of the subfamilies contained Platy-1 elements perfectly matching their respective consensus

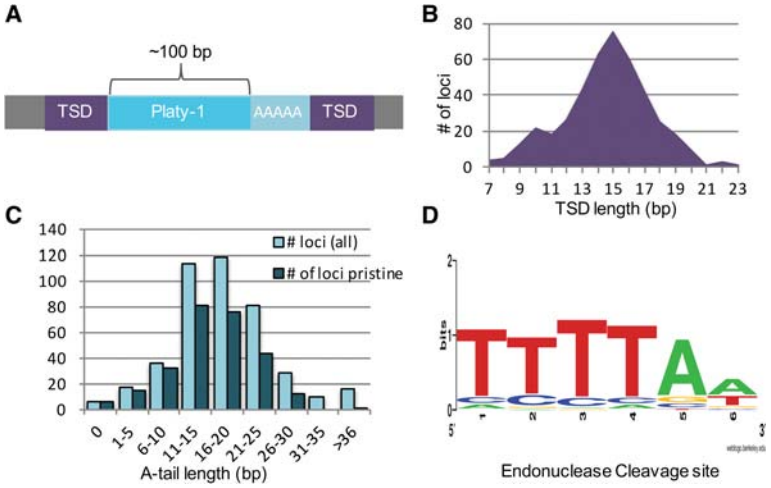

Figure 1. Platy-1 characteristics. (*A*) The structure of Platy-1 (turquoise). The element terminates in an A-tail (light turquoise) and is flanked by TSDs (purple). Flanking sequence is shown in dark gray. (*B*) The TSD distribution length across 424 elements. (*C*) The A-tail length distribution of all 424 Platy-1 elements as well as the length distribution of pristine A-tails. (*D*) The endonuclease cleavage site across all elements with TSDs is illustrated as a Weblogo (Crooks et al. 2004).

Table 1. Platy-1 subfamily sizes and age estimates

subfamily	length	#/subfamily	min % div	max % div	avg % div	age[1] (my)	age[2] (my)	age[3] (my)	age[4] (my)
1	104	55	4.9	26.8	14.7	66.4	24.5	267.8	73.2
2	104	27	3.9	26.4	12.3	55.4	20.4	223.5	61.1
2a	104	51	3.9	16	10.5	47.2	17.4	190.4	52.0
2b	104	35	7.8	27.6	13.7	61.5	22.7	248.4	67.9
3	104	20	4.9	16.5	9.89	44.5	16.4	179.8	49.1
4	104	19	1.9	9.4	4.92	22.2	8.2	89.5	24.4
4a	104	19	5	14.8	8.07	36.4	13.4	146.7	40.1
5	108	8	2.8	19.3	7.87	35.5	13.1	143.1	39.1
6	104	15	1	25	5.87	26.4	9.7	106.7	29.2
6a	104	74	1	21.2	4.66	21.0	7.7	84.7	23.2
6b	104	80	1	9.7	4.31	19.4	7.2	78.4	21.4
6c	104	129	1	13.7	4.61	20.8	7.7	83.8	22.9
6d	104	12	1	5.8	3.35	15.1	5.6	60.9	16.6
6e	104	14	1.9	7.8	4.65	20.9	7.7	84.5	23.1
6f	104	27	0	10.9	5.8	26.1	9.6	105.5	28.8
6g	104	32	1	12.6	5.19	23.4	8.6	94.4	25.8
6h	104	61	0	10.8	4.44	20.0	7.4	80.7	22.1
6x	104	32	0	9.9	4.54	20.5	7.5	82.5	22.6
7	104	66	0	7.9	3.36	15.1	5.6	61.1	16.7
7a	104	17	0	6.9	3.46	15.6	5.7	62.9	17.2
8	104	25	2	10.8	5.27	23.7	8.7	95.8	26.2
9	104	37	0	7.9	3	13.5	5.0	54.5	14.9
9a	104	30	0	7.8	3.7	16.7	6.1	67.3	18.4
9b	104	97	0	6.8	2.32	10.5	3.9	42.2	11.5
9c	104	45	0	6.7	2.84	12.8	4.7	51.6	14.1
9d	104	33	0	6.8	3.04	13.7	5.0	55.3	15.1
9e	104	13	0	5.9	2.32	10.5	3.9	42.2	11.5
10	104	43	0	9.7	3.41	15.4	5.7	62.0	16.9
10a	104	18	0	6.8	3.14	14.1	5.2	57.1	15.6
11	104	35	0	6.7	2.52	11.4	4.2	45.8	12.5
11a	104	16	0	6.8	3.23	14.5	5.4	58.7	16.0
11b	104	21	1	9.9	3.54	15.9	5.9	64.4	17.6
11c	104	33	1	9.7	3.06	13.8	5.1	55.6	15.2
12	104	19	0	4.8	2.2	9.9	3.7	40.0	10.9
12a	104	12	0	5.9	2.83	12.7	4.7	51.5	14.1
12b	104	55	0	2.9	1.24	5.6	2.1	22.5	6.2
12c	104	18	0	4.8	2.11	9.5	3.5	38.4	10.5
12d	104	19	0	4.8	1.64	7.4	2.7	29.8	8.1
12e	104	20	0	3.9	1.7	7.7	2.8	30.9	8.4
12f	104	21	0	6	1.95	8.8	3.2	35.5	9.7
13	104	159	0	12.6	1.31	5.9	2.2	23.8	6.5
13a	104	15	0	5.8	1.42	6.4	2.4	25.8	7.1
13b	104	13	0	2.9	1.22	5.5	2.0	22.2	6.1
13c	104	35	0	9.7	2.3	10.4	3.8	41.8	11.4
13d	104	25	0	5.8	1.96	8.8	3.3	35.6	9.7
13e	104	63	0	4.8	1.83	8.2	3.0	33.3	9.1
13f	104	30	0	4.8	1.94	8.7	3.2	35.3	9.6
13g	104	43	0	4.8	1.65	7.4	2.7	30.0	8.2
14	95	143	0	7.4	1.99	9.0	3.3	36.2	9.9
14a	95	20	0	4.3	1.46	6.6	2.4	26.5	7.3
14b	95	16	0	5.4	2.29	10.3	3.8	41.6	11.4
15	105	95	0	9.8	1.63	7.3	2.7	29.6	8.1
15a	105	16	0	5.8	1.76	7.9	2.9	32.0	8.7
16	104	42	0	4.8	1.78	8.0	3.0	32.4	8.8
16a	104	20	0	4.9	2.14	9.6	3.6	38.9	10.6
16b	104	23	0	2.9	1.23	5.5	2.0	22.4	6.1
16c	104	7	0	2	0.84	3.8	1.4	15.3	4.2
16d	104	10	0	1.9	1.06	4.8	1.8	19.3	5.3
16e	104	8	0	5	1.12	5.0	1.9	20.4	5.6
16f	104	51	0	6.9	2.09	9.4	3.5	38.0	10.4
17	82	20	0	6.3	3.79	17.1	6.3	68.9	18.8
17a	82	11	2.5	9.1	6.05	27.3	10.0	110.0	30.1

The age was calculated assuming two different mutation rates. For the calculation, the average divergence of each subfamily was used. Gray data bars indicate relative values with respect to size or divergence estimates. (#) number; (my) million years ago; (age[1]) assumed mutation rate of 2.2×10^{-9} (Kumar and Subramanian 2002); (age[2]) assumed mutation rate of 7.53×10^{-10} (Perez et al. 2013); (age[3]) assumed mutation rate of 0.55×10^{-9} (Lipson et al. 2015), integrating 29 yr/generation human-referenced estimate; (age[4]) assumed mutation rate of 0.55×10^{-9} (Lipson et al. 2015), 6 yr generation time (Perez et al. 2013). These numbers should be regarded as rough estimates.

sequence. Intriguingly, 26 subfamilies had three or more identical members, with one subfamily having 44 perfect Platy-1 elements. Altogether, almost 10% (224/2268) of Platy-1 elements are pristine, and more than a quarter (628/2268) have a divergence of 1.5% or less (Fig. 2). Our analyses show that both the average percent divergence and the divergence spectrum (i.e., element with lowest/highest substitution rate) varied considerably between subfamilies, indicating that some subfamilies were active for a longer time period than others (Table 1). Alternately, subfamilies with a wider spectrum may harbor more than one subfamily that cannot be distinguished due to high divergence and/or low number of elements.

Since non-LTR retrotransposons mutate at a neutral rate (Cordaux et al. 2006), the divergence from the consensus sequence can be used to approximate the age of a subfamily if the mutation rate is known and the molecular clock is constant. However, the molecular clock appears heterogeneous across primates (Li and Tanimura 1987; Hwang and Green 2004; Steiper et al. 2004; Kim et al. 2006; Steiper and Young 2006; Perelman et al. 2011), making time estimates more complicated. For mammals, an average neutral substitution rate of 2.2×10^{-9} per base per year has been suggested (Kumar and Subramanian 2002), whereas human neutral substitution rates based on pedigrees have recently been estimated to be between 1 and 1.2×10^{-8} per base per generation (Roach et al.

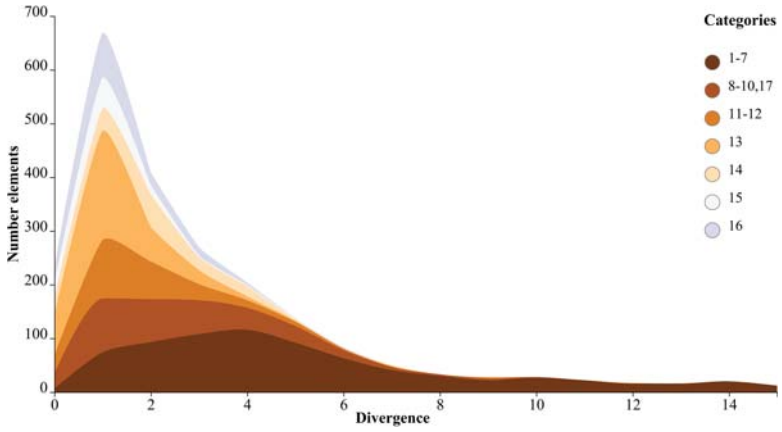

Figure 2. Platy-1 evolution in NWMs. The histogram shows the Platy-1 distribution based on the divergence from the consensus sequence of all 2268 full-length sequences. The subfamilies are color-coded based on subfamily affiliation and grouped together based on age. The divergence from the respective consensus sequence was retrieved from RepeatMasker and is shown on the *x*-axis. The *y*-axis shows the number of elements with the indicated divergence. The plot is generated with custom BioPython scripts and the Vega + D3 Vincent wrapper/package (Bostock et al. 2011) (http://github.com/wrobstory/vincent).

2010; Conrad et al. 2011; Kong et al. 2012; Scally and Durbin 2012). Alternate approaches utilize a comparison of current-day genomes and precisely dated ancient genomes with estimate ranges from 1.1 to 1.7×10^{-8} per base per generation (Fenner 2005; Fu et al. 2014). A recent study utilizes the fine-scale human recombination map for mutation rate calibration, resulting in a mutation rate of 1.6×10^{-8} per base per generation or $0.55 \pm 0.05 \times 10^{-9}$ per base per year (Lipson et al. 2015).

Non-CpG SNPs have been linked to the generation time and body size of species (Kim et al. 2006; Perez et al. 2013), which can vary considerably within primates. While a convergent slowdown in primate sequence evolution rates has been proposed (Steiper and Seiffert 2012), this could not be confirmed for platyrrhines (Perez et al. 2013) in part due to varying body size and generation time within NWMs. Given that the neutral mutation rate within and between species continues to be a subject of debate, we provide average age estimates for Platy-1 subfamilies assuming different mutation rates (Table 1). The divergence spectrum of Platy-1 elements (Fig. 2) suggests a slow propagation rate with only a few subfamilies active in early NWM evolution. That is further supported by neighbor-joining (Saitou and Nei 1987; Kuhner and Felsenstein 1994) and median-joining network analyses (Fig. 3; Methods; Supplemental Methods S2; Bandelt et al. 1999).

Platy-1 genomic distribution

Our analysis of Platy-1 elements shows a scattered distribution within and across all chromosomes, excluding insertions on "chromosome" Random (Fig. 4; Supplemental Fig. S3A). All chromosomes except Y contain Platy-1 insertions; this is not unexpected, as the Y Chromosome is relatively small in size, and the genome of a female marmoset contains only a small fraction (3.5 Mb) of Y chromosomal DNA from her fraternal twin due to chimerism. The intrachromosomal distribution appears to be random, supported by a random insertion model (Supplemental Methods S3; Supplemental Fig. S3B). In contrast, we rejected a random insertion model for Platy-1 distribution across the genome (i.e., the number of insertions is not proportional to the size of the chromosome) based on the results of a χ^2 analysis ($\chi^2 = 51.35$, df = 25, $P = 0.000656$). Hence, Platy-1 elements are not evenly distributed across all chromosomes. Chromosome 4 exhibited a much lower density and Chromosomes 18, 19, and 22 a higher density (Fig. 4). An uneven distribution of non-LTR retrotransposons across the chromosomes has been observed previously (e.g., SVA or *Alu* elements) (International Human Genome Sequencing Consortium 2001; Wang et al. 2005).

Origin of the Platy-1 repeat family

Next, we investigated the origin of the Platy-1 repeat family. The internal promoter regions of known SINEs seem to be derived from tRNA, 7SL RNA, or 5S RNA. Another alternative, though less likely, is the origin of a mobile element through horizontal transfer (Gilbert et al. 2012). To confirm that Platy-1 was not derived from tRNA, we performed a tRNA-Scan (Lowe and Eddy 1997; Schattner et al. 2005) and confirmed absence of a cloverleaf secondary structure using Mfold (Zuker 2003). Our BLAST (Altschul et al. 1990) query of the whole database (excluding *C. jacchus)* with consensus sequences of the oldest Platy-1 subfamilies identified mostly matches in primates, including human, indicating that Platy-1 may be derived from 7SL RNA. Our multiple sequence alignment of the Platy-1 consensus sequences against 7SL RNA, FRAM, FLAM, and a selection of *Alu*J/S consensus sequences further supports that Platy-1 is derived from 7SL RNA (Fig. 5; Supplemental Fig. S4). More specifically, Platy-1 may be derived from an *Alu*J element based on sequence identity.

Amplification dynamics of Platy-1 repeat family

In order to determine when Platy-1 elements started amplifying in primates, we compared full-length Platy-1 candidate loci with orthologous sequences retrieved from the human genome (hg19), which represents the genome with the highest quality of all primate genomes. We did not find convincing evidence of Platy-1 elements shared between common marmoset and human. Given that some of the oldest elements may have accumulated mutations to a degree that they are no longer recognized as Platy-1-derived, we next screened the human genome for the presence of Platy-1 elements through a RepeatMasker analysis, revealing 27 full-length Platy-1 candidate loci (Methods; Supplemental Table S4). We discarded 14 loci based on manual analysis because of lack of TSDs and high sequence similarity upstream of or downstream from the putative Platy-1 sequence, with other Platy-1 candidate loci, an indication for nonproper Platy-1 insertions after ruling out duplication events.

The remaining 13 putative Platy-1 loci appear to have been active early in Platy-1 evolution based on the average sequence divergence and absence of additional Platy-1 insertions in the rhesus macaque genome (rheMac3) (Supplemental Table S5). For 10 of the catarrhine-specific Platy-1 candidate loci, we were able to identify TSDs, A-tails of varying length, and endonuclease cleavage sites (Supplemental Table S6), suggesting insertion through retrotransposition. Two additional loci likely represent Platy-1 insertions; however, these insertions occurred into adenosine-rich sequences, obscuring accurate TSD identification. For one putative Platy-1 locus (Chr 16: 74696055–74696158 [hg19]), we could not

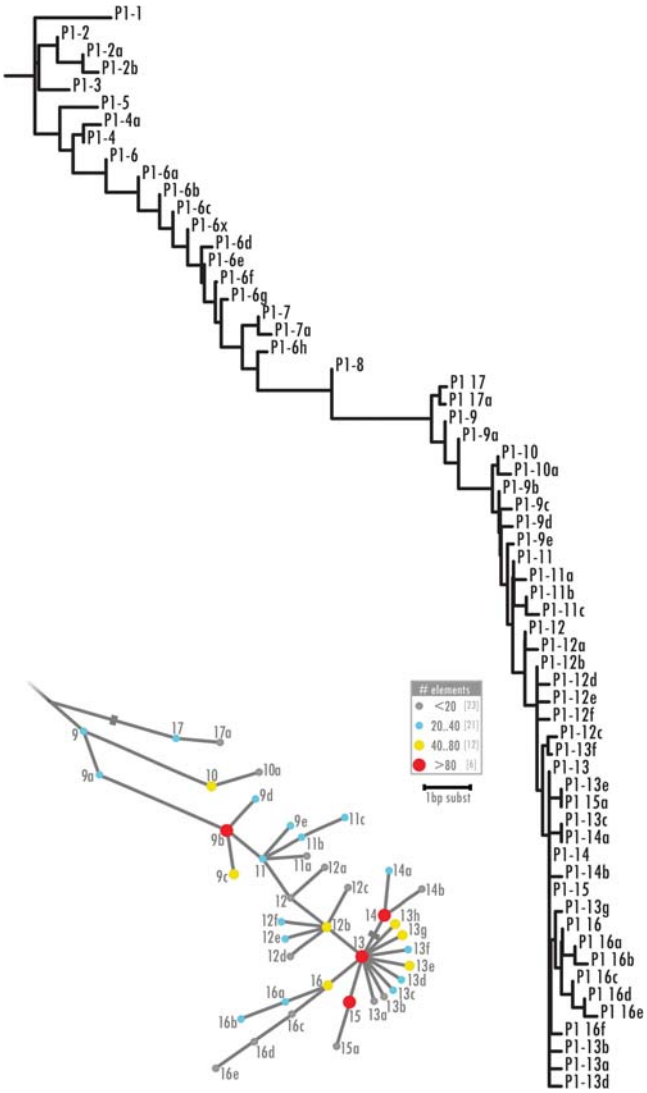

Figure 3. Platy-1 subfamily tree reconstruction. A neighbor-joining tree for all 62 Platy-1 subfamilies is shown. The excerpt (*bottom left*) shows a network analysis for younger subfamilies. The nodes for each subfamily represent the approximate size of each subfamily based on the number of full-length Platy-1 elements.

identify TSDs. This locus terminated in an A-tail immediately followed by a stretch of thymine nucleotides. The majority (seven of 13) of the putative Platy-1 loci were absent from the common marmoset genome. Four loci, including the locus without TSDs, were shared with common marmoset; none of these were present in our common marmoset data set. For two loci, we could not unambiguously determine lineage specificity due to (partial) absence of the flanking sequence in *C. jacchus*. Conceivably, the Platy-1 candidate locus on Chr 16 (hg19) could be the founding sequence for Platy-1 because it is present in both catarrhines and platyrrhines, and terminates in a pristine A-tail immediately followed by a Pol III termination signal.

Our computational analysis of the Platy-1 repeat structure indicates the presence of older subfamilies as well as subfamilies of more recent origin. To better elucidate the amplification dynamics of Platy-1 elements, we selected 308 full-length Platy-1 candidate loci across the divergence spectrum for our PCR-based phylogenetic analyses. We analyzed 271 loci on our phylogenetic panel (Fig. 6; Supplemental Tables S1A, S6) and received informative results for 210 loci. Altogether, 101 putative Platy-1 loci were excluded based on no/unspecific amplification (37 loci), missing amplification of more than four NWM species (10 loci), no amplification of the closest related species with resolved phylogenetic relationships (50 loci altogether; most commonly Platy-1 amplicon in marmosets and no amplification in tamarin), or no amplification of Platy-1 in any species (four loci).

Occasionally, we encountered amplicon sizes of unexpected length (i.e., a size different from the predicted filled or empty amplicon size) in at least one NWM species. Given the different amplicon size in one or more species, these loci were easily identified. Intriguingly, we identified one locus (on Chr 13, locus 157) for which the amplicon pattern, i.e., a PCR product of the same approximate size as the amplicon containing a Platy-1 element, suggested a different relationship than the majority of the tested loci, which were in agreement with previous phylogenies (Ray et al. 2005; Osterholz et al. 2009). More specifically, the insertion suggested a close relationship between marmosets and the family Atelidae. DNA sequencing of this locus revealed a truncated *Alu* insertion of similar length as the Platy-1 insertion (Supplemental Fig. S5; Supplemental Information S8).

All Platy-1 insertions were exclusive to platyrrhines. A few Platy-1 insertions (10 loci) were shared across all NWM families, and the majority (60.95%) of the sampled Platy-1 loci were specific to marmosets, with a smaller number of elements (38 loci) being unique to common marmoset (Fig. 6). In addition, we identified 31 Callithrichinae- and one Cebidae-specific Platy-1 insertions. Our analyses of the common marmoset-specific insertions on the population panel of 24 common marmosets (Supplemental Table S1A) show that some (three loci) Platy-1 elements are polymorphic within *C. jacchus* (Supplemental Table S7).

Discussion

We identified a new SINE family, Platy-1, in the common marmoset genome that is most likely derived from *AluJ*. A full-length Platy-1 element is only ~100 bp in length, making it the shortest SINE in primates. A few active Platy-1 subfamilies contain deletions, illustrating that even shorter SINEs can maintain retrotranspositional activity. Although the A box shows considerable variation during the evolution of Platy-1, the B box has primarily remained stable. This, taken together with the presence of TSDs, an endonuclease cleavage site, and termination in an A-tail without polyadenylation signal indicates transcription of Platy-1 by Pol III and insertion into the genome via TPRT by the enzymatic machinery of L1. The lack of a polyadenylation signal prior to the A-tail also indicates that the A-tail is derived from the source element as previously reported for *Alu* elements (Batzer et al. 1990; Deininger and Batzer 1993; Shaikh and Deininger 1996). The absence of TSDs for ~10% of Platy-1 elements is likely primarily caused by decay due to the age of some of the insertions. Alternate reasons include insertion in sequences with simple repeat characteristics and insertion of other mobile elements into the tail immediately upstream of or downstream from Platy-1. A fraction of these insertions may also have occurred through endonuclease-independent insertion mechanisms as described for *Alu* and L1 elements (Morrish et al. 2002; Gilbert et al. 2005; Sen et al. 2007; Srikanta et al. 2009).

A-tail length and no accumulation of other nucleotides in their tails have been associated with higher retrotransposition

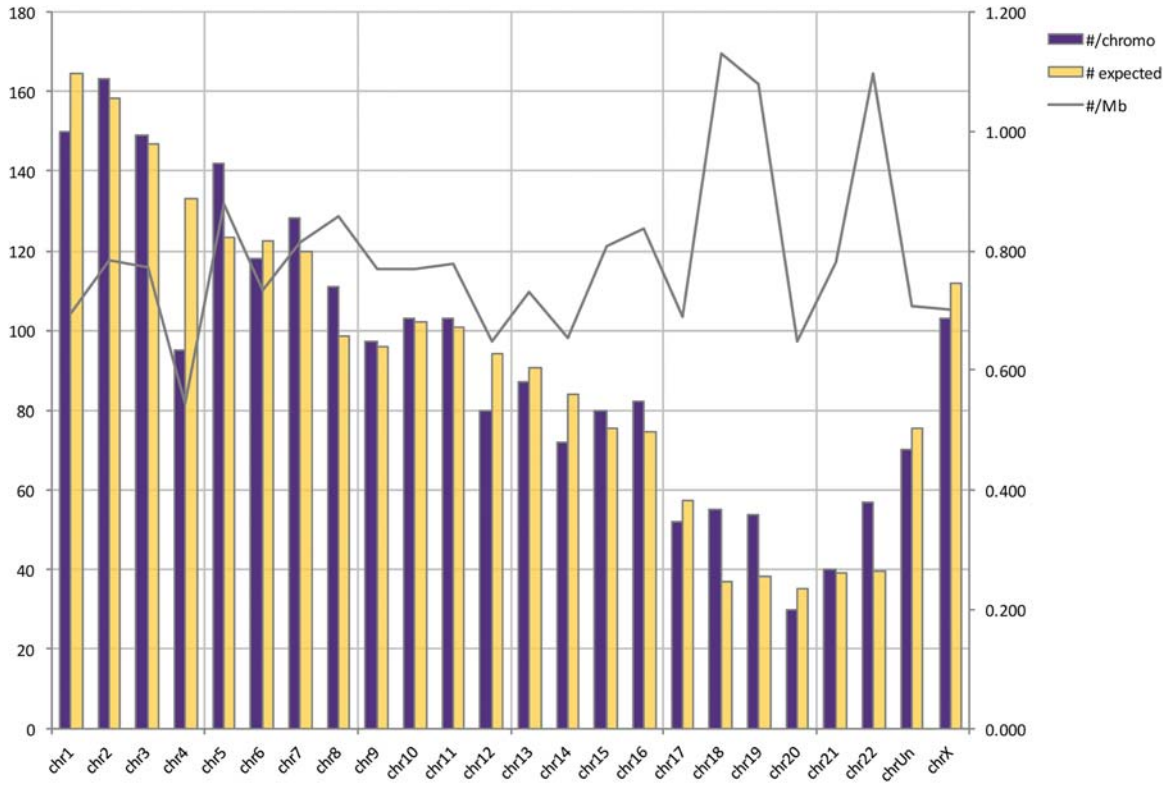

Figure 4. Platy-1 genomic distribution. The expected (yellow) and actual (purple) Platy-1 distributions across all chromosomes (excluding Chr Random and Chr Y) are illustrated. In addition, the density per megabase is shown for each chromosome. Due to omission of putative Platy-1 loci on Chr Random, 2221 full-length elements were included in this analysis.

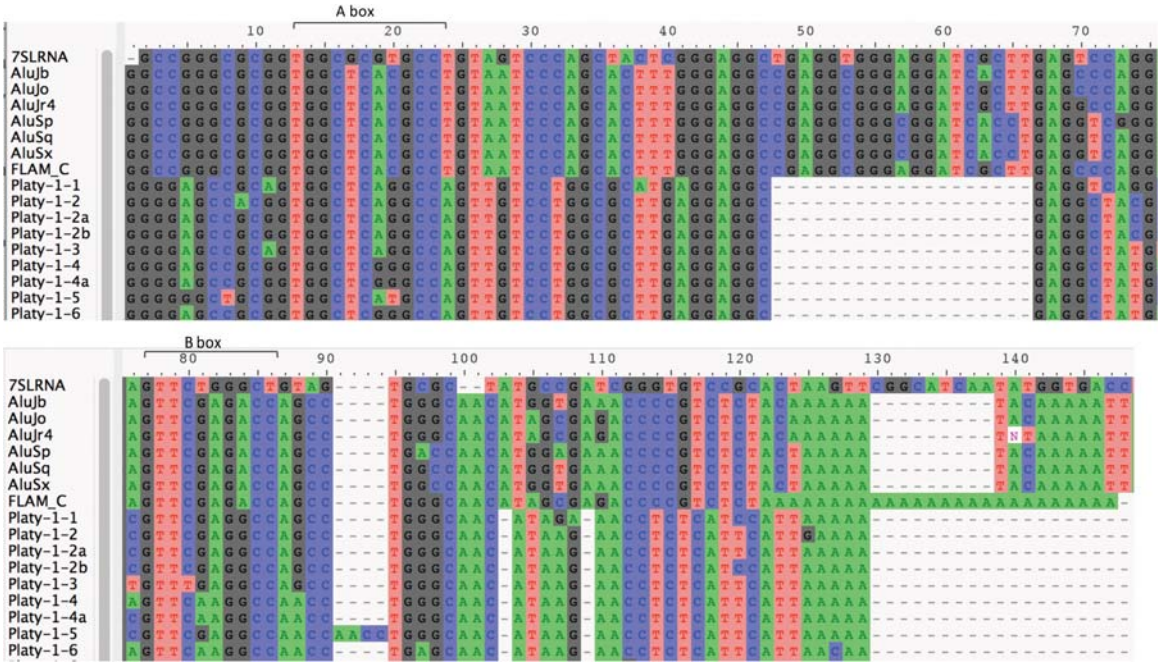

Figure 5. Alignment of Platy-1 with 7SL RNA and *Alu* elements. Shown is a multiple sequence alignment using MUSCLE (Edgar 2004) followed by manual curation of the oldest Platy-1 subfamilies with 7SL RNA, a selection of *Alu* consensus sequences, and FLAM. The alignment is visualized with AliView (Larsson 2014). Dashes indicate absence of the sequence. Also illustrated are the A box (consensus sequence: TRGYnnAnnnG) and B box (consensus sequence: GWTCRAnnCc). The tail of the Platy-1 sequence aligned equally well to the regions prior to the middle A-rich region and the 3′ end of an *Alu* element. In the latter case, the deletion may have been caused by recombination between homologous sequences. This alignment assumes that the element terminates at the middle A-rich region. An alternate alignment is provided in Supplemental Figure S4.

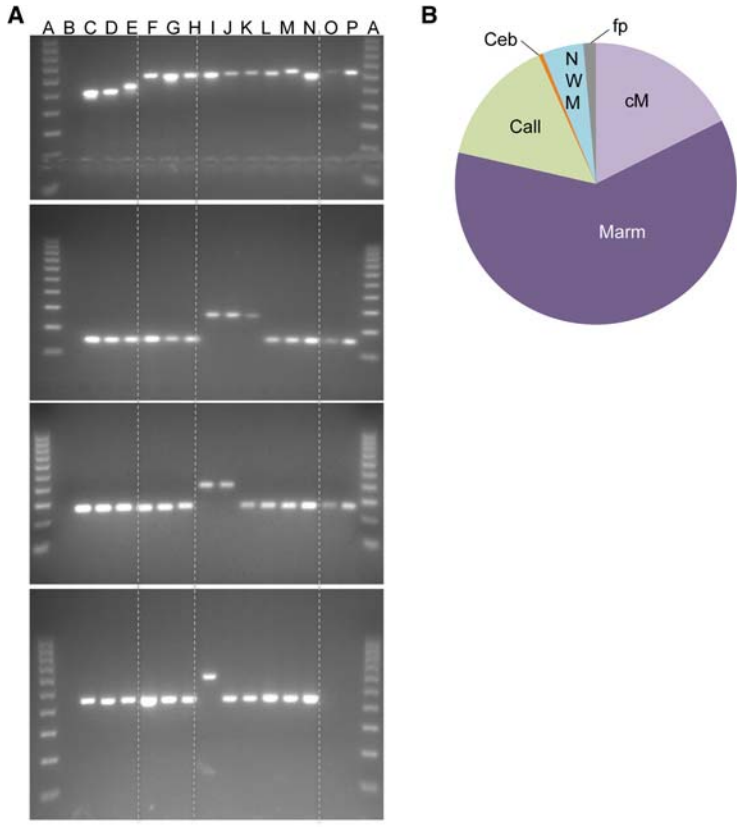

Figure 6. Phylogenetic distribution of Platy-1. (*A*) Four agarose gel chromatographs of our locus-specific phylogenetic analyses. An *upper* fragment indicates presence of a Platy-1 insertion; a *lower* fragment, absence. The vertical lines from *left* to *right* separate the outgroups from NWMs, Atelidae from Cebidae, and Cebidae from Pitheciidae. The gel chromatographs show (from *left* to *right*): (A) 100 bp ladder; (B) TLE; (C) human; (D) common chimpanzee; (E) African green monkey; (F) woolly monkey; (G) spider monkey; (H) red howler monkey; (I) common marmoset; (J) pygmy marmoset; (K) tamarin; (L) capuchin monkey; (M) squirrel monkey; (N) owl monkey; (O) titi; (P) saki. (For more detailed information regarding the species used, please see Supplemental Table S1A.) The *top* gel image shows a Platy-1 insertion shared across all NWMs. The gel chromatograph *below* shows an insertion specific to Callithrichinae, which is followed by a marmoset-specific insertion. The gel chromatograph on the *bottom* shows a common marmoset-specific Platy-1 insertion. (*B*) The phylogenetic results for our informative loci are shown in a pie chart: (Marm) marmosets; (cM) common marmoset; (fp) false positive; (NWM) New World monkey; (Call) Callithrichinae; (Ceb) Cebidae.

activity of *Alu* (Roy-Engel et al. 2002; Bennett et al. 2008; Comeaux et al. 2009); this is likely true for Platy-1 as well. Based on this and the analysis of 427 elements, we estimate that ~8% of elements have a higher chance of being retrotranspositionally active (pristine A-tail and length >20 bp). Of these, more than half contained a Pol III termination signal downstream within 100 bp and 28% within 25 bp. The latter elements have the highest probability of being source elements, as a Pol III termination signal immediately downstream from an *Alu* element has been positively correlated with *Alu* retrotransposition efficiency in vitro (Comeaux et al. 2009). On the other hand, more than one-third of the inspected Platy-1 A-tails harbored nucleotide substitutions, with many of them containing microsatellites. This is in agreement with previous findings that A-tails of non-LTR retrotransposons, particularly *Alu* elements, represent seeds for microsatellite formation (Arcot et al. 1995; Nadir et al. 1996; Fungtammasan et al. 2012; Grandi and An 2013; Grandi et al. 2013) and indicates a role of Platy-1 in the birth of microsatellites, and consequently, to structural variation beyond insertional mutagenesis.

Lineage-specific repeat families containing *Alu* sequence have been identified in some primate lineages. For example, the hominoid-specific SVA element (Ostertag et al. 2003; Wang et al. 2005; Damert et al. 2009; Hancks and Kazazian 2010) and the gibbon-specific LAVA element (Carbone et al. 2012) harbor *Alu* sequence as a component of these composite mobile elements. In the *Galago crassicaudatus*, a lineage-specific SINE derived from an *Alu* element, into which a Type III element inserted, has been reported (Daniels and Deininger 1985; Roos et al. 2004). That said, Platy-1 not only represents the first lineage-specific SINE in NWMs, but also the first SINE in primates that is solely derived from an *Alu* element.

Platy-I propagation

Overall, our results support the birth of Platy-1 around the time of divergence of catarrhines and platyrrhines, as both lineages show at least some evidence for Platy-1 mobilization. However, Platy-1 established itself as a repeat family only in NWMs. This raises the question why this may be the case, especially since the potential founding element for the Platy-1 repeat family, an element with a perfect A-tail followed immediately by a Pol III termination signal, appears to be present in both lineages. It may be that this sequence lost its propagation properties in the lineage leading to human prior to the rise of daughter elements with propagation capabilities. Alternatively, modifications in the upstream sequence (e.g., nucleotide substitutions or sequence rearrangements) may have played a role, as it is known that the upstream sequence of *Alu* elements is crucial for its retrotransposition capabilities (Roy et al. 2000b). However, the exact upstream sequence requirements remain elusive.

Platy-1 provided the unique opportunity to study its propagation dynamics from its birth up to the present—spanning more than 40 million years of evolution. The mammalian neutral substitution rate seems to better reconstruct the evolution of Platy-1 (Table 1; Supplemental Table S4) with respect to primate radiation estimates. This does not necessarily imply that the mammalian substitution rate is more accurate, as the relatively short length of the Platy-1 element does not allow for a fine-scale divergence resolution. Also, its high GC content, resulting in above-average CpG sites, may contribute to the uncertainty of the average substitution rate and (consequently) age estimate of Platy-1 subfamilies. Moreover, the reconstruction of older subfamilies is less accurate due to mutation accumulation. Gene conversion represents another factor that would result in a higher divergence of Platy-1 sequences and as a result could impact age estimates of subfamilies. However, we believe that gene conversion plays a limited role for

Platy-1 elements given the lower number of these elements in the genome and the short length of Platy-1; both factors generally reduce the risk of gene conversion. Moreover, gene conversion events should be relatively easily detected, and these elements would have been grouped as a separate subfamily with hybrid sequence characteristics during manual curation, similar to those found in human *Alu* gene conversion events (Kass et al. 1995; Roy et al. 2000a).

Following an initially slow mobilization rate, Platy-1 propagation seems to have continuously increased until propagation reached a peak that coincided with the rise of the marmoset ancestor, suggesting a recent Platy-1 expansion with several retrotransposition-competent subfamilies propagating in parallel. This is in stark contrast to the early Platy-1 evolution, in which propagation was primarily linear (Fig. 3) and more closely resembled the typical subfamily evolution of LINEs. This is likely a result of the low number of source elements early in Platy-1 evolution—initially, of a single subfamily. In contrast, the more recent history of Platy-1 follows a star-like pattern, typical for the propagation of *Alu* elements (Cordaux et al. 2006). Most recently, Platy-1 retrotransposition has somewhat slowed based on the relatively small number of polymorphic Platy-1 insertions within common marmoset populations (Supplemental Table S7). However, Platy-1 has been active very recently and is likely still propagating in common marmosets.

The timing of the recent Platy-1 expansion differs from the *Alu* propagation dynamics in the lineage leading to the common marmoset (The Marmoset Genome Sequencing and Analysis Consortium 2014). The history of *Alu* elements shows an exceptional retrotransposition peak about 40 million years ago (International Human Genome Sequencing Consortium 2001; The Marmoset Genome Sequencing and Analysis Consortium 2014), coinciding with the estimated divergence of the NWM lineage from the ancestral anthropoid lineage about 35–47 million years ago. This was followed by a decline of the *Alu* propagation rate both in Platyrrhini (The Marmoset Genome Sequencing and Analysis Consortium 2014) and Catarrhini (International Human Genome Sequencing Consortium 2001). A second smaller *Alu* retrotransposition peak occurred in the NWM lineage, leading to the common marmoset in more recent history, and has subsequently declined (The Marmoset Genome Sequencing and Analysis Consortium 2014). Based on the divergence estimates, Platy-1 propagation peaked more recently than *Alu* retrotransposition in the NWM lineage and coincided with the rise of the common ancestor of marmosets. The more recent amplification success of Platy-1 may be the result of different factors, including a higher number of source elements in conjunction with escape from the host response against Platy-1 elements. Although speculative, the decline of the *Alu* mobilization rate could at least partially be caused by the higher number of retrotransposition-competent Platy-1 elements, as both Platy-1 and *Alu* likely compete for the enzymatic machinery of L1. In general, the interplay of different factors, including number of retrotransposition-competent elements, preference of the L1 enzymatic machinery, and host factors, all affect the amplification dynamics of non-LTR retrotransposons and result in varying retrotransposition rates over time.

Taken together, we identified and characterized a new SINE in the NWM lineage leading to the common marmoset. The birth of Platy-1 coincided with the rise of NWMs, allowing for the investigation of the Platy-1 propagation dynamics throughout the radiation of NWMs. The current study was performed through the lens of the common marmoset. It remains to future studies to determine the fate of Platy-1 in other NWM species.

Methods

Analysis of the original Platy-1 element

We retrieved orthologous nucleotide sequences for the original multispecies sequence alignment on Chromosome 3 using BLAT (Kent 2002) from the common marmoset (calJac3.2), human (hg19), chimpanzee (panTro2/panTro4), orangutan (ponAbe2), and rhesus macaque (rheMac2/rheMac3) genome assemblies. Assisted by the ClustalW function of BioEdit (Hall 1999), we performed a multiple species alignment. Next, we queried the original Platy-1 sequence against the *C. jacchus* genome (calJac3.2) using BLAT, as well as against all other aforementioned genome assemblies and the squirrel monkey draft genome assembly (saiBol1). We also queried the original sequence against the whole database using BLASTN (Altschul et al. 1990) to determine if the sequence was identified in species other than the common marmoset. To assess if the sequence was derived from a known repeat, we checked the sequence with RepeatMasker (Smit et al. 2013–2015).

Subfamily reconstruction

We selected the 10 best matches from our *C. jacchus* BLAT query and reconstructed a preliminary query sequence using a majority rule approach. We included this sequence in the library of our local RepeatMasker analysis (Supplemental Methods S1) and retrieved all full-length Platy-1 hits, defined as start position <4 bp and end position not shorter than two nucleotides prior to the A-tail. Following multiple sequence alignments with ClustalW (Hall 1999) and/or MUSCLE (Edgar 2004), and together with manual curation, the Platy-1 elements were sorted based on presence/absence of SNPs and short indels (deletions/insertions of nucleotides) using a majority rule approach (for further information, please refer to Supplemental Methods S1). Following the creation of Platy-1 consensus sequences, we RepeatMasked the *C. jacchus* assembly and determined for each subfamily (1) if the minimum requirements for the presence of a subfamily was met (i.e., minimum number of elements); and (2) if we could identify additional subfamilies. This step was repeated until we identified and confirmed all subfamilies.

Phylogenetic subfamily tree construction

We reconstructed the subfamily evolution using DNAdist, neighbor, and drawgram from the Phylip v3.695 (Felsenstein 1989) analytics suite for generation of a neighbor-joining tree using the Kimura-2-parameter model. As input, we used a file containing all Platy-1 subfamilies. We manually moved subfamilies with deletions to the appropriate locations in the tree because these subfamilies evolved most likely from each other, as the independent recurrence of the same deletion is highly unlikely. In parallel, we performed a network analysis using Network version 4.612 (Supplemental Methods S2; Bandelt et al. 1999). To avoid misplacing subfamilies, we removed all deletions except for one nucleotide from the respective consensus sequences.

Platy-1 sequence feature analysis

For our Platy-1 sequence analyses, we scrutinized the elements as well as sequences immediately upstream and downstream for the presence of common sequence features. Specifically, A-tail length and composition, TSDs, and endonuclease cleavage sites were manually determined and recorded. For the determination of

TSDs, we allowed for the presence of up to four SNPs, because TSDs decay over time—likely at a neutral rate similar to non-LTR retrotransposons (Cordaux et al. 2009). We excluded loci without TSDs from our downstream analyses and used Weblogo (Crooks et al. 2004) to visualize the endonuclease cleavage site.

Platy-l genomic distribution

We calculated the density of full-length Platy-1 elements across the *C. jacchus* genome by counting Platy-1 elements per chromosome divided by the number of megabases of nucleotide sequences for each chromosome (excluding ChrRand, which, reduced the data set from 2268 to 2221 putative Platy-1 loci) and performed a χ^2 analysis. We utilized the GATK software (McKenna et al. 2010) to calculate the GC content of the flanking sequence. We also determined the intrachromosomal Platy-1 distribution and performed simulations of a random intrachromosomal distribution model (Supplemental Methods S3).

Determination of Platy-l lineage specificity

To determine when Platy-1 originated, we queried all full-length Platy-1 elements, including 500 bp of flanking sequence, against the human genome (hg19) using a local BLAT installation (Kent 2002) and retrieved the resulting sequences. We aligned the human sequence against the *C. jacchus* sequence using MUSCLE (Edgar 2004) for matches in human of <10,000 bp in length and manually determined the presence of Platy-1 in catarrhines. In addition, we RepeatMasked the human genome (hg38) with a custom library and retrieved all full-length Platy-1 elements, including 500 bp of flanking sequence. Next, we performed the same analysis as described above by querying against the common marmoset genome and checked all putative Platy-1 loci for the presence of typical non-LTR features. We also RepeatMasked the rhesus macaque (rheMac3) genome, retrieved all full-length Platy-1 insertions, and compared Platy-1 insertions from the rhesus macaque genome with hits from the human genome.

Age estimate calculation

Based on neutral mutation rate estimates of 8.5×10^{-10}–6.06×10^{-10} per base per generation for crown Platyrrhini and of 9.07×10^{-10}–6.47×10^{-10} for crown Cebidae (Perez et al. 2013), we assumed an average substitution rate of 7.53×10^{-10} per base per generation for our subfamily age estimates. We used 8 yr as an approximate generation time given that marmosets have a generation time of about 6 yr, and the generation likely was longer earlier in primate radiation. We also calculated the average age of the elements using an estimated neutral substitution rate for mammals of 2.2×10^{-9} per year per base (Kumar and Subramanian 2002) and one derived from human studies using 0.55×10^{-8} and varying generation times (Lipson et al. 2015).

Oligonucleotide primer design and PCR analyses

For primer design, we retrieved the Platy-1 element plus 600 bp of flanking on either site from the *C. jacchus* genome using BLAT (Kent 2002). Orthologous sequences to the flanking sequence were retrieved from the human (hg19), chimpanzee (panTro2/4), orangutan (ponAbe3), and rhesus macaque (rheMac2/3) genome assemblies using BLAT; and a multiple species alignment was performed with ClustalW (Hall 1999). We designed primers with Primer3 (Rozen and Skaletsky 2000) using the default settings with the following exceptions: only mononucleotide repeats with up to four consecutive identical nucleotides were permitted, the minimum nucleotide number for primers was increased to 20, the annealing temperature range was set to 57°C–61°C, and the allowed PCR amplicon size was increased to 1400 bp. All primer pairs that passed our downstream screening (Supplemental Methods S4) were ordered from Sigma Aldrich (Supplemental Table S8).

In preparation for sequencing (Supplemental Methods S5) and wet bench-based phylogenetic analyses (for species and origin of DNA, see Supplemental Table S1A), we performed PCRs in a 96-well format using a BioRad iCycler thermocycler in a final volume of 25 µL. Each PCR reaction contained 25 ng of template DNA; 200 nM of each oligonucleotide primer; 1.5 mM $MgCl_2$; 1X PCR buffer (50 mM KCl; 10 mM TrisHCl, pH 8.3); 0.2 mM dNTPs; and 2 units *Taq* DNA polymerase. PCR reactions were performed using the following conditions: initial denaturation for 90 sec at 94°C, followed by 32 cycles of denaturation for 30 sec at 94°C, annealing at optimal temperature for 20 sec, and extension for 30 sec at 72°C. PCRs were terminated with a final extension for 3 min at 72°C. The PCR amplicons (20 µL of each PCR product) were size fractionated in a horizontal gel chamber on a 2% agarose gel containing 0.1 µg/mL ethidium bromide for 50 min at 200 V. DNA fragments were visualized with UV-fluorescence and images were saved using a BioRad ChemiDoc XRS imaging system.

Data access

Sequencing data generated for this study have been submitted to GenBank (http://www.ncbi.nlm.nih.gov/genbank/) under accession numbers KT427526–28 and are shown in the alignment of Supplemental Figure S5.

Acknowledgments

We thank Jerilyn A. Walker and Catherine C. Fontenot for discussion, support, and advice. This research was supported by the National Institutes of Health R01 GM59290 (M.A.B.). E.L.A. and S.S. were supported in part by a grant to Louisiana State University from the Howard Hughes Medical Institute (HHMI) through the Precollege and Undergraduate Science Education Program. S.S. was also supported in part by the Louisiana Board of Regents Supervised Undergraduate Research Experience (SURE) program (B.U., M.K.K., M.A.B.). The computational analyses were in part supported through National Science Foundation grant CNS-1126739 (B.U., M.A.B., M.K.K.). Portions of this research were conducted with high-performance computing resources provided by LONI and LSU HPC.

Author contributions: M.K.K. found the element, designed the studies, performed analyses, and wrote the manuscript. B.U. performed computational analyses and was involved in the writing of the manuscript; E.L.A. and S.S. performed phylogenetic and population genetic analyses; E.L.A., S.S., and S.A.B. were involved in Platy-1 characterization. R.H. and A.F.A.S. were involved in the identification and characterization of the element. M.A.B. provided experimental input, support, and helped write the manuscript.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA. 1995. *Alu* repeats: a source for the genesis of primate microsatellites. *Genomics* **29:** 136–144.

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16:** 37–48.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370–379.

Batzer MA, Kilroy GE, Richard PE, Shaikh TH, Desselle TD, Hoppens CL, Deininger PL. 1990. Structure and variability of recently inserted Alu family members. *Nucleic Acids Res* **18:** 6793–6798.

Belancio VP, Hedges DJ, Deininger P. 2008. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res* **18:** 343–358.
Bennett E, Keller H, Mills R, Schmidt S, Moran J, Weichenrieder O, Devine S. 2008. Active *Alu* retrotransposons in the human genome. *Genome Res* **18:** 1875.
Bostock M, Ogievetsky V, Heer J. 2011. D^3 data-driven documents. *IEEE Trans Vis Comput Graph* **17:** 2301–2309.
Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci* **100:** 5280–5285.
Burns KH, Boeke JD. 2012. Human transposon tectonics. *Cell* **149:** 740–752.
Carbone L, Harris RA, Mootnick AR, Milosavljevic A, Martin DI, Rocchi M, Capozzi O, Archidiacono N, Konkel MK, Walker JA, et al. 2012. Centromere remodeling in *Hoolock leuconedys* (Hylobatidae) by a new transposable element unique to the gibbons. *Genome Biol Evol* **4:** 648–658.
Carbone L, Harris RA, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513:** 195–201.
Chen JM, Stenson PD, Cooper DN, Férec C. 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum Genet* **117:** 411–427.
The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.
Comeaux MS, Roy-Engel AM, Hedges DJ, Deininger PL. 2009. Diverse *cis* factors controlling *Alu* retrotransposition: what causes *Alu* elements to die? *Genome Res* **19:** 545–555.
Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, et al. 2011. Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43:** 712–714.
Cordaux R, Lee J, Dinoso L, Batzer MA. 2006. Recently integrated *Alu* retrotransposons are essentially neutral residents of the human genome. *Gene* **373:** 138–144.
Cordaux R, Batzer M, Box P. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10:** 691–703.
Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188–1190.
Damert A, Raiz J, Horn AV, Löwer J, Wang H, Xing J, Batzer MA, Löwer R, Schumann GG. 2009. 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res* **19:** 1992–2008.
Daniels GR, Deininger PL. 1985. Repeat sequence families derived from mammalian tRNA genes. *Nature* **317:** 819–822.
Deininger PL, Batzer MA. 1993. Evolution of retroposons. *Evol Biol* **27:** 157–196.
Deininger PL, Batzer MA, Hutchison CA III, Edgell MH. 1992. Master genes in mammalian repetitive DNA amplification. *Trends Genet* **8:** 307–311.
Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35:** 41–48.
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5:** 113.
Felsenstein J. 1989. PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* **5:** 164–166.
Feng Q, Moran JV, Kazazian HH Jr, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87:** 905–916.
Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol* **128:** 415–423.
Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PL, Aximu-Petri A, Prüfer K, de Filippo C, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514:** 445–449.
Fungtammasan A, Walsh E, Chiaromonte F, Eckert KA, Makova KD. 2012. A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? *Genome Res* **22:** 993–1005.
Gilbert N, Lutz S, Morrish TA, Moran JV. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25:** 7780–7795.
Gilbert C, Hernandez SS, Flores-Benabib J, Smith EN, Feschotte C. 2012. Rampant horizontal transfer of *SPIN* transposons in squamate reptiles. *Mol Biol Evol* **29:** 503–515.
Grandi FC, An W. 2013. Non-LTR retrotransposons and microsatellites: partners in genomic variation. *Mob Genet Elements* **3:** e25674.
Grandi FC, Rosser JM, An W. 2013. LINE-1-derived poly(A) microsatellites undergo rapid shortening and create somatic and germline mosaicism in mice. *Mol Biol Evol* **30:** 503–512.
Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41:** 95–98.
Han K, Xing J, Wang H, Hedges DJ, Garber RK, Cordaux R, Batzer MA. 2005. Under the genomic radar: the stealth model of *Alu* amplification. *Genome Res* **15:** 655–664.
Hancks DC, Kazazian HH Jr. 2010. SVA retrotransposons: evolution and genetic instability. *Semin Cancer Biol* **20:** 234–245.
Huang CR, Burns KH, Boeke JD. 2012. Active transposition in genomes. *Annu Rev Genet* **46:** 651–675.
Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci* **101:** 13994–14001.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Kass DH, Batzer MA, Deininger PL. 1995. Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Mol Cell Biol* **15:** 19–25.
Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res* **12:** 656–664.
Kim SH, Elango N, Warden C, Vigoda E, Yi SV. 2006. Heterogeneous genomic molecular clocks in primates. *PLoS Genet* **2:** e163.
Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488:** 471–475.
Konkel MK, Batzer MA. 2010. A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* **20:** 211–221.
Konkel MK, Walker JA, Batzer MA. 2010. LINEs and SINEs of primate evolution. *Evol Anthropol* **19:** 236–249.
Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* **11:** 459–468.
Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci* **99:** 803–808.
Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30:** 3276–3278.
Li WH, Tanimura M. 1987. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326:** 93–96.
Lipson M, Loh PR, Sankararaman S, Patterson N, Berger B, Reich D. 2015. Calibrating the human mutation rate via ancestral recombination density in diploid genomes. *PLoS Genet* **11:** e1005550.
Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang SP, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orangutan genomes. *Nature* **469:** 529–533.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25:** 955–964.
Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72:** 595–605.
The Marmoset Genome Sequencing and Analysis Consortium. 2014. The common marmoset genome provides insight into primate biology and evolution. *Nat Genet* **46:** 850–857.
Mathias SL, Scott AF, Kazazian HH Jr, Boeke JD, Gabriel A. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254:** 1808–1810.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303.
Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* **31:** 159–165.
Nadir E, Margalit H, Gallily T, Ben-Sasson SA. 1996. Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. *Proc Natl Acad Sci* **93:** 6470–6475.
Osterholz M, Walter L, Roos C. 2009. Retropositional events consolidate the branching order among New World monkey genera. *Mol Phylogenet Evol* **50:** 507–513.
Ostertag EM, Kazazian HH Jr. 2001. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* **35:** 501–538.
Ostertag EM, Goodier JL, Zhang Y, Kazazian HH Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73:** 1444–1451.
Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular phylogeny of living primates. *PLoS Genet* **7:** e1001342.

Perez SI, Tejedor MF, Novo NM, Aristide L. 2013. Divergence times and the evolutionary radiation of New World monkeys (Platyrrhini, Primates): an analysis of fossil and molecular data. *PLoS One* **8:** e68029.
Ray DA, Xing J, Hedges DJ, Hall MA, Laborde ME, Anders BA, White BR, Stoilova N, Fowlkes JD, Landry KE, et al. 2005. *Alu* insertion loci and platyrrhine primate phylogeny. *Mol Phylogenet Evol* **35:** 117–126.
Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316:** 222–234.
Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328:** 636–639.
Roos C, Schmitz J, Zischler H. 2004. Primate jumping genes elucidate strepsirrhine phylogeny. *Proc Natl Acad Sci* **101:** 10650–10654.
Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AO, Batzer MA, Deininger PL. 2000a. Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res* **10:** 1485–1495.
Roy AM, West NC, Rao A, Adhikari P, Aleman C, Barnes AP, Deininger PL. 2000b. Upstream flanking sequences and transcription of SINEs. *J Mol Biol* **302:** 17–25.
Roy-Engel AM, Salem AH, Oyeniran OO, Deininger L, Hedges DJ, Kilroy GE, Batzer MA, Deininger PL. 2002. Active Alu element "A-tails": size does matter. *Genome Res* **12:** 1333–1344.
Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132:** 365–386.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406–425.
Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* **13:** 745–753.
Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33:** W686–W689.
Schrago CG, Russo CA. 2003. Timing the origin of New World monkeys. *Mol Biol Evol* **20:** 1620–1625.
Sen SK, Huang CT, Han K, Batzer MA. 2007. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res* **35:** 3741–3751.
Shaikh TH, Deininger PL. 1996. The role and amplification of the HS Alu subfamily founder gene. *J Mol Evol* **42:** 15–21.
Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9:** 657–663.
Smit A, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.
Srikanta D, Sen SK, Huang CT, Conlin EM, Rhodes RM, Batzer MA. 2009. An alternative pathway for *Alu* retrotransposition suggests a role in DNA double-strand break repair. *Genomics* **93:** 205–212.
Steiper ME, Seiffert ER. 2012. Evidence for a convergent slowdown in primate molecular rates and its implications for the timing of early primate evolution. *Proc Natl Acad Sci* **109:** 6006–6011.
Steiper ME, Young NM. 2006. Primate molecular divergence dates. *Mol Phylogenet Evol* **41:** 384–394.
Steiper ME, Young NM, Sukarna TY. 2004. Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid–cercopithecoid divergence. *Proc Natl Acad Sci* **101:** 17021–17026.
Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3:** research0052.
Walker JA, Konkel MK, Ullmer B, Monceaux CP, Ryder OA, Hubley R, Smit AF, Batzer MA. 2012. Orangutan Alu quiescence reveals possible source element: support for ancient backseat drivers. *Mob DNA* **3:** 8.
Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: a hominid-specific retroposon family. *J Mol Biol* **354:** 994–1007.
Zingler N, Willhoeft U, Brose HP, Schoder V, Jahns T, Hanschmann KM, Morrish TA, Löwer J, Schumann GG. 2005. Analysis of 5′ junctions of human LINE-1 and *Alu* retrotransposons suggests an alternative model for 5′-end attachment requiring microhomology-mediated end-joining. *Genome Res* **15:** 780–789.
Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415.

Received September 3, 2015; accepted in revised form February 23, 2016.

Discovery of a new repeat family in the *Callithrix jacchus* genome

Miriam K. Konkel, Brygg Ullmer, Erika L. Arceneaux, et al.

Genome Res. 2016 26: 649-659 originally published online February 25, 2016
Access the most recent version at doi:10.1101/gr.199075.115